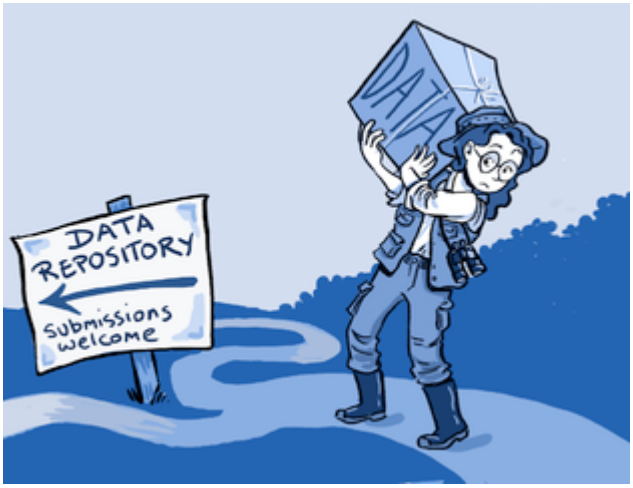
Digital Data Realities

written by Jeremy Huggett | 29/06/2016



The Cost of Digital Data (Ainsley Seago via Wikimedia Commons) CC BY 4.0

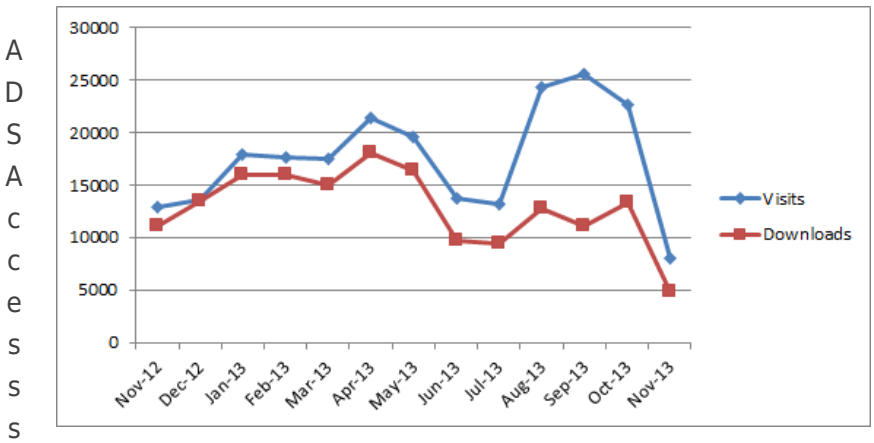
The UK is suddenly wakening from the reality distortion field that has been created by politicians on both sides and only now beginning to appreciate the consequences of Brexit – our imminent departure from the European Union. But – without forcing the metaphor – are we operating within some kind of archaeological reality distortion field in relation to digital data?

Undoubtedly one of the big successes of digital archaeology in recent years has been the development of digital data repositories and, correspondingly, increased access to archaeological information. Here in the UK we've been fortunate enough to have seen this develop over the past twenty years in the shape of the Archaeology Data Service, which offers search tools, access to digital back-issues of journals, monograph series and grey literature reports, and the availability of downloadable datasets from a variety of field and research projects. In the past, large-scale syntheses took years to complete (for instance, Richard Bradley's synthesis of British and Irish prehistory took four years paid research leave with three years of research assistant support in order to travel the country to seek out grey literature reports accumulated over 20 years (Bradley 2006, 10)). At this moment, there are almost 38,000 such reports in the Archaeology Data Service digital library, with more are added each month (a more than five-fold increase since January 2011, for example). The appearance of projects of synthesis such as the Rural Settlement of Roman Britain is starting to provide evidence of the value of access to such online digital resources. And, of course, other countries increasingly have their own equivalents of the ADS – tDAR and OpenContext in the USA, DANS in the Netherlands, and the Hungarian National Museum's Archaeology Database, for instance).

But all is not as rosy in the archaeological digital data world as it might be.

Earlier this month, a number of the original pioneers of the Internet and the World Wide Web joined together to call for a decentralized Web (Perry 2016), arguing that it has become a collection of silos rather than its original decentralised vision as a result of the short-term gains brought by centralisation and the compromises this creates. This strikes an archaeological chord as it rather seems as if we've been busy creating a series of data silos both on a national and international scale with little if any communication between them. Indeed, the original vision for the ADS in 1996 was explicitly decentralised, consisting of interlinked resources brokered through a centralised interface. At the time it was perhaps overly ambitious, and it's not that this hasn't been tried – there were pioneering attempts around the turn of the millennium by the ADS. For example, the ARENA project (Kenny & Richards 2005) sought to link together various European archaeological data resources, but is no longer available; similarly the ADS's HEIRPORT (Historic EnvIRronment PORTal) allowed a search to be conducted simultaneously across a number of heritage databases within the UK (Fernie 2003), but is again no longer functional. The Transatlantic Archaeology Gateway (TAG) is available and demonstrates what is possible, enabling searches across both the ADS and tDAR. However, incorporation of datasets derived from Canmore and other online resources into the ADS system is still achieved through taking snapshots, rather than live links, and there is little evidence of interlinked systems elsewhere. Current European projects, such as the European Research Infrastructure for Heritage Science (E-RIHS), suggest that further attempts at interlinking will be made, but otherwise the strong impression of a series of siloed resources remains.

Providing the hooks to allow others to build systems linking to existing resources is one approach to overcoming this. The ADS provides a limited testbed of Linked Open Data created through the STELLAR project but specifically does not provide any support or advice in its use. tDAR doesn't advertise it has an API but if you dig around its documentation, one is available. OpenContext is by far the most up-front about its API and provides the most accessible information; it's also recently announced a competition to encourage its use. However, there is still more than a sense of black art about linked open data: the situation is broadly similar to that described by Matthew Lincoln (2016) recently. Following his trials and tribulations of using linked open data for research, one of his recommendations is for usable interfaces to LOD repositories, and it seems that the problems encountered in the early 2000s with unreliable connections (mmm … memories of Z39.50!) to linked resources haven't gone away.



tatistics (from Huggett 2015)

But whoever may or may not be creating the links between our separate repositories, the fact

remains that – whisper it quietly – few of us are actually using the archived digital data in the first place. It'll be interesting to follow the progress of the OpenContext competition in this respect. ADS access statistics show a surprising number of downloads relative to visits, but this disguises that the majority of downloads are of PDFs of past issues of journal articles, and especially grey literature reports, rather than data. So the reality seems to be that that these repositories, whilst popular and heavily visited, are seeing little use of their carefully curated datasets. Is this a problem with the way the data are presented? Is it simply too soon in the cycle, because we don't normally revisit such data until a generation or so has passed? Is it down to a lack of good examplars of the re-use of this kind of data (for instance, the promised detailed information about the ADS Digital Data Reuse Award winners/highly commended entries was never blogged)? Or is it something else altogether?

Fortunately, other aspects and facilities offered by data repositories are used and demonstrably valued – for instance, a study of the ADS calculated a direct use value of the ADS of around £1.4m per annum relative to an annual investment of £1.2m, but beyond that the efficiency impacts were estimated at anything between £13m and £58m per annum (Beagrie and Houghton 2013). That's remarkable – but the sad fact remains that actual income for repositories such as these is far from secure. The ADS was funded by the AHRC for many years but no longer – it primarily relies on research and commercial income (for instance, see its costing calculator). Similarly, repositories such as tDAR and the Hungarian National Museum Archaeology Database are funded by grant income for now, though tDAR intends to transform into a not-for-profit organisation and advertises its digital preservation costs for depositors. OpenContext is supported via an impressive constellation of sponsors and puts its deposit charges on its front page but overall it seems none of the archaeological repositories have a secure income stream, despite their crucial archival role.

So despite the success in developing and maintaining digital data repositories at not inconsiderable cost financially and in personal efforts, we find ourselves in a triple bind where we have a set of resources which are as yet underlinked, underused in certain key respects, and insecurely funded. Arguably addressing the first two may go some way to approaching the last of these but it remains to be seen.

None of this negates the importance of what has been achieved. But in our valid and valuable pursuit of developing digital data resources, it's all too easy to overlook some fairly fundamental issues. For example, dana boyd has, in the context of computer coding, recently asked four simple questions that should give pause for thought:

> "Does the system that we built produce the right output given the known constraints? Do we understand the biases and limitations of the system and the output? Are those clear to the user so that our tool cannot enable poor decision-making or inaccurate impressions? What are the true social and environmental costs of the service?" (boyd 2016).

The answer to the first question is *'maybe'* – it all depends on what these constraints are and whether we really know or understand them. The second answer is *'no'* – we don't fully understand or appreciate the biases and limitations and no amount of metadata, paradata, or whateverdata we might provide can hope to disentangle this as it is not purely a technical problem but is in the nature of archaeological data. Nor are the systems we are building transparent in terms of how they

operate. The third answer is also *'no'*, as we can't mitigate against incorrect or mistaken use of our data once they are online – even if we are careful about explaining things like this, we can't assume a user will take any notice. The fourth question is not something that is frequently considered, if at all. But as dana boyd says:

> "… is the world really better off having petabytes of data sitting on live servers just to make sure it's open and accessible just in case? It's painful to think about how many terabytes of data are sitting in open data repositories that have never been accessed." (boyd 2016).

And then there's the inherent bias within our repositories. Although most contain data drawn from projects undertaken elsewhere around the world, the archaeological community is not especially diverse. Tim Hitchcock has recently argued that the transfer of public analogue archives into the digital realm has had the effect of making Western data and values hyper-available and flattened out the range and diversity of human experience (Hitchcock 2016). To what extent is this equally true of archaeological repositories and their data? Should we not seek to facilitate more, non-western, archaeological repositories (and link to them!), which is not the same as making their data available in our repositories?

A reality distortion field is not necessarily a bad thing if the outcomes are positive. After all, it is "a phenomenon in which an individual's intellectual abilities, persuasion skills and persistence make other people believe in the possibility of achieving very difficult tasks". We need to continue to use our collective intellectual abilities, powers of persuasion, and persistence to meet the challenges presented by the demands of access, presentation, and sustainability of archaeological digital data. As Ted Nelson famously said: "Everything is deeply intertwingled", and until we incorporate this in our infrastructure designs, we will barely scratch the surface of what is possible.

(Some of these thoughts were in part sparked by the Expert Forum on the *Future of Archaeological Knowledge Curation 2021-2026* organised by the ARIADNE SIG on Archaeological Research Practices and Methods hosted by the Digital Curation Unit-IMIS, Athena Research Centre – my thanks in particular to Costis Dallas, Agiatis Benardou and Nephelie Chatzidiakou and their colleagues).

References

Beagrie, N. and Houghton, J. 2013 *The value and impact of the Archaeology Data Service: A study and methods for enhancing sustainability*. http://archaeologydataservice.ac.uk/research/impact

boyd, d. 2016 'Be Careful What You Code For' *Points: Data & Society,* https://points.datasociety.net/be-careful-what-you-code-for-c8e9f3f6f55e#

Bradley, R. 2006 'Bridging the Two Cultures – Commercial Archaeology and the Study of Prehistoric Britain', *The Antiquaries Journal* 86, 1–13.

Fernie, K. 2003 'Getting it together on-line: HEIRNET and Internet-based resource discovery tools for the Historic Environment', *Internet Archaeology* 13. http://intarch.ac.uk/journal/issue13/fernie_index.html

Hitchcock, T. 2016 ‘Privatising the Digital Past’, *Historyonics*, http://historyonics.blogspot.co.uk/2016/06/privatising-digital-past.html

Huggett, J. 2015 ‘Digital haystacks: open data and the transformation of archaeological knowledge’. In: Wilson, A. T. and Edwards, B. (eds.) *Open Source Archaeology: Ethics and Practice.* De Gruyter Open, pp. 6-29. http://eprints.gla.ac.uk/114652/

Kenny, J. and Richards, J. 2005 ‘Pathways to a Shared European Information Infrastructure for Cultural Heritage’, *Internet Archaeology* 18. http://intarch.ac.uk/journal/issue18/kenny_index.html

Lincoln, M. 2016 ‘Linked Open Realities: The Joys and Pains of Using LOD for Research’ http://matthewlincoln.net/2016/06/06/linked-open-realities-the-joys-and-pains-of-using-lod-for-research.html

Perry, T. 2016 ‘The Fathers of the Internet Revolution Urge Today’s Software Engineers to Reinvent the Web’, *IEEE Spectrum* http://spectrum.ieee.org/view-from-the-valley/telecom/internet/the-fathers-of-the-internet-revolution-urge-todays-pioneers-to-reinvent-the-web